Functional Analytic Theory of Supervised Learning (Common Ground between Functional Analysis and Mathematical Theory of Information)

AUTHOR(S):

Hirabayashi, Akira; Ogawa Hidemitsu

数理解析研究所講究録 1253 巻 2002 年 1-13

関数解析的学習理論

Functional Analytic Theory of Supervised Learning

山口大学・工学部　平林　晃（Hirabayashi Akira）
Faculty of Engineering, Yamaguchi University
東京工業大学・大学院 情報理工学研究科　小川　英光（Ogawa Hidemitsu）
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology

Abstract

A new horizon has been opened for functional analysis. Supervised learning is a fruitful problem for functional analysis. This paper reviews a functional analytic theory of supervised learning. It is a problem of estimating an underlying rule of training samples. When the rule can be expressed by a function, supervised learning can be regarded as a function approximation problem. We first formulate the problem by using a reproducing kernel Hilbert space. Then, three kinds of learning, i.e., projection learning, partial projection learning, and averaged projection learning are introduced. They are unified into the concept of a family of projection learnings, which includes an infinite kind of learning. By using the general theory of the family, minimum variance projection learning is introduced.

1 Introduction

Learning is one of the most important abilities of human beings. For example, most children can speak and understand language by the time they are about five years old. By imitating what adults are doing, they seem to obtain some rule for speaking or understanding language.

As shown in this example, obtaining an underlying rule by using samples is supervised learning. When the rule can be expressed by a function, supervised learning can be regarded as a function approximation problem [7]. Hence, it is naturally connected to functional analysis.

In this paper, we review a functional analytic theory of supervised learning. First, the problem of supervised learning is formulated by using a reproducing kernel Hilbert space. Then, projection learning [7], partial projection learning [9], and averaged projection

learning [13] are introduced. Furthermore, these individual theories of learning are unified into a family of projection learnings. It is a theory that treats an infinite kind of learning in a unified way. By using the theory, minimum variance projection learning is introduced.

2 Learning as an Inverse Problem

In this section, a framework to discuss the learning problem proposed in [5, 7] is reviewed. Supervised learning is a problem of estimating an underlying rule of training examples. In this paper, we discuss the case that the rule can be expressed by a complex-valued function. Let $f(\boldsymbol{x})$ be a target function where $\boldsymbol{x}$ is an L-dimensional vector. We assume that f belongs to a Hilbert space, H, which has the reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}')$. Let $\mathcal{D}$ be the domain of functions in H.

The reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$ that satisfies the following two conditions [1]:

- For any fixed $\boldsymbol{x}'$ in $\mathcal{D}$, $K(\boldsymbol{x}, \boldsymbol{x}')$ belongs to H as a function of $\boldsymbol{x}$.
- It holds for any f in H and any $\boldsymbol{x}'$ in $\mathcal{D}$ that

$$\langle f(\cdot), K(\cdot, \boldsymbol{x}') \rangle = f(\boldsymbol{x}'), \tag{1}$$

  where $\langle \cdot, \cdot \rangle$ is the inner product in H.

In a general Hilbert space, a function f is treated as a point. Hence, a value of f at a point $\boldsymbol{x}$, denoted by $f(\boldsymbol{x})$, can not be discussed. If the space has the the reproducing kernel, however, $f(\boldsymbol{x})$ can be discussed by using the inner product of the space as shown in Eq.(1). In supervised learning, it is an essential requirement for functions to have a value at a point $\boldsymbol{x}$ as shown in Eq.(2) below.

The goal of learning is to obtain a function, f_0, which is the best approximation to f under some learning criterion J using training examples $\{\boldsymbol{x}_m, y_m\}_{m=1}^M$, where

$$y_m = f(\boldsymbol{x}_m) + n_m, \tag{2}$$

and n_m is additive noise. Hence, supervised learning can be considered as a function approximation problem.

This problem can be formulated by using a functional analytic approach. Let $\boldsymbol{y}$ and $\boldsymbol{n}$ be vectors in an M-dimensional unitary space $\mathbf{C}^M$ whose m-th elements are y_m and n_m, respectively. Once $\{\boldsymbol{x}_m\}_{m=1}^M$ is fixed, $\{f(\boldsymbol{x}_m)\}_{m=1}^M$ is uniquely determined from f. Therefore, we can introduce a sampling operator, A, which maps f to the vector consisting of $\{f(\boldsymbol{x}_m)\}_{m=1}^M$:

$$Af = \begin{pmatrix} f(\boldsymbol{x}_1) \\ \vdots \\ f(\boldsymbol{x}_M) \end{pmatrix}. \tag{3}$$

It can be expressed by the reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ and the Neumann-Schatten product [10] as

$$A = \sum_{m=1}^{M} \boldsymbol{e}_m \otimes \overline{K(\cdot, \boldsymbol{x}_m)}, \tag{4}$$

where $\{\boldsymbol{e}_m\}_{m=1}^{M}$ are standard basis in $\mathbf{C}^M$, i.e., the M-dimensional vector consisting of zero elements except the m-th element equal to 1. Note that A is linear even when we are concerned with nonlinear functions.

The relationship between f and $\boldsymbol{y}$ can be expressed by

$$\boldsymbol{y} = Af + \boldsymbol{n}. \tag{5}$$

Let X be an operator which maps $\boldsymbol{y}$ to f_0:

$$f_0 = X\boldsymbol{y}. \tag{6}$$

X is called a learning operator. Eqs.(5) and (6) reformulates the learning problem as that of obtaining X which provides the best approximation f_0 to f from $\boldsymbol{y}$ under the criterion J. We discuss the case that X is linear.

One of the most widely used learning criteria is the training error defined by

$$J_M[X] = \sum_{m=1}^{M} |f_0(\boldsymbol{x}_m) - y_m|^2 \tag{7}$$

$$= \|AX\boldsymbol{y} - \boldsymbol{y}\|^2, \tag{8}$$

where $\|\cdot\|$ denotes the norm in $\mathbf{C}^M$. J_M evaluates only the errors between y_m and $f_0(\boldsymbol{x}_m)$ $(m = 1, 2, \ldots, M)$. Our goal is, however, to minimize an overall error between f and f_0.

The following criterion is also used widely:

$$J_R[X] = J_M[X] + \alpha\|T(X\boldsymbol{y})\|^2, \tag{9}$$

where α is a positive constant, and T is a linear operator. For example, T is the identity operator or a differential operator. $\|T(X\boldsymbol{y})\|^2$ in Eq.(9) is called a regularization term, and the technique to add the term to J_M is called regularization. J_R still does not evaluate the overall error between f and f_0.

In order to evaluate the error between f and f_0, one of the author, H. Ogawa, has proposed projection learning, partial projection learning, and averaged projection learning, which are described in the following section.

3 Projection Learning, Partial Projection Learning, and Averaged Projection Learning

We start with projection learning, because it create partial projection learning, averaged projection learning, and so on. Let A^* be the adjoint operator of A, $\mathcal{R}(A^*)$ be the range of A^*, and $P_{\mathcal{R}(A^*)}$ be the orthogonal projection operator onto $\mathcal{R}(A^*)$.

Definition 1 (Projection learning [4, 6]) *An operator X is called a projection learning operator and denoted by $A^{(\mathrm{P})}$ if X, under the constraint*

$$XA = P_{\mathcal{R}(A^*)}, \tag{10}$$

minimizes the following functional

$$J_n[X] = E_{\boldsymbol{n}}\|X\boldsymbol{n}\|^2, \tag{11}$$

where $E_{\boldsymbol{n}}$ is the ensemble average for $\{\boldsymbol{n}\}$ and $\|\cdot\|$ is the norm in H. Obtaining f_0 by an $A^{(\mathrm{P})}$ is called 'projection learning'.

Projection learning is based on the following idea. Eqs.(5) and (6) yield

$$f_0 = XAf + X\boldsymbol{n}. \tag{12}$$

The first and second terms in the right-hand side of Eq.(12) are called the signal component and the noise component of f_0, respectively. The target function f is unknown. However, we discuss the case that the space H to which f belongs is known. Hence, we attempt to achieve that signal component XAf agrees with the best approximation for each f in H. The totality of XAf for all f in H is the subspace $\mathcal{R}(XA)$, and the best approximation of f in $\mathcal{R}(XA)$ is the orthogonal projection of f onto this subspace. Therefore, we try to achieve for each f in H that

$$XAf = P_{\mathcal{R}(XA)}f, \tag{13}$$

which yields

$$XA = P_{\mathcal{R}(XA)}. \tag{14}$$

It follows from Eq.(14) that $\mathcal{R}(XA) = \mathcal{R}((XA)^*) = \mathcal{R}(A^*X^*) \subset \mathcal{R}(A^*)$. Hence, it holds that

$$\mathcal{R}(XA) \subset \mathcal{R}(A^*). \tag{15}$$

The larger $\mathcal{R}(XA)$ provides the better approximation $P_{\mathcal{R}(XA)}f$. Hence, we consider the largest $\mathcal{R}(XA)$, i.e.,

$$\mathcal{R}(XA) = \mathcal{R}(A^*). \tag{16}$$

Eqs.(14) and (16) yield Eq.(10).

Eqs.(10) and (12) imply that f_0 is distributed around $P_{\mathcal{R}(A^*)}f$ due to $X\boldsymbol{n}$. Hence, the variance of the distribution is minimized, which is expressed by $J_n[X]$ in Eq.(11). These discussions lead us to Definition 1.

In partial projection learning and averaged projection learning, the noise component is suppressed in the same way as projection learning. That is, the functional $J_n[X]$ is minimized under some constraint for the signal component, such as Eq.(10). The constraints make difference between these three projection learnings.

Partial projection learning is used for the case that a target function f belongs to a closed subspace S in H.

Definition 2 (Partial projection learning [9]) *An operator X is called a partial projection learning operator and denoted by $A^{(\mathrm{PTP})}$ if X minimizes Eq.(11) under the constraint*

$$XAP_S = P_{\mathcal{R}(P_S A^*)} P_S, \tag{17}$$

where P_S and $P_{\mathcal{R}(P_S A^*)}$ are the orthogonal projection operators onto S and $\mathcal{R}(P_S A^*)$, respectively. Obtaining f_0 by an $A^{(\mathrm{PTP})}$ is called 'partial projection learning.'

In case of $S = H$, the constraint Eq.(17) reduces to Eq.(10). In other words, projection learning is a special case of partial projection learning. Note that $A^{(\mathrm{PTP})}A$ is the orthogonal projection operator onto $\mathcal{R}(P_S A^*)$ along $S \cap \mathcal{N}(A)$ when A is restricted to S, where $\mathcal{N}(A)$ is the null space of A.

Averaged projection learning is used for the case that more accurate learning results are required for functions that appear more frequently.

Definition 3 (Averaged projection learning [13]) *An operator X is called an averaged projection learning operator and denoted by $A^{(\mathrm{AP})}$ if X minimizes Eq.(11) under the constraint that X minimizes*

$$J_{\mathrm{AP}}[X] = E_f \|XAf - f\|^2, \tag{18}$$

where E_f is the ensemble average with respect to $\{f\}$. Obtaining f_0 by an $A^{(\mathrm{AP})}$ is called 'averaged projection learning.'

The constraint Eq.(18) seems to be different from constraints in projection learning and partial projection learning. The averaged projection learning operator, however, has the following property as well as projection learning and partial projection learning. From Eq.(18), we have the following normal equation:

$$XARA = RA, \tag{19}$$

where R is the correlation operator of the ensemble $\{f\}$. Let $\overline{\mathcal{R}(R)}$ be the closure of $\mathcal{R}(R)$. Eq.(19) implies that the operator $A^{(\mathrm{AP})}A$ is the projection operator onto $\mathcal{R}(RA^*)$ along $\overline{\mathcal{R}(R)} \cap \mathcal{N}(A)$ when it is restricted to $\overline{\mathcal{R}(R)}$ [14].

Properties of projection learning, partial projection learning, and averaged projection learning have been studied in detail in, for example, [4, 8, 14]. They can be discussed in a unified way by a theory of a family of projection learnings which is described in the following section.

4 A Family of Projection Learnings

In this section, individual theories of learning are unified into a family of projection learnings. It is a theory that treats an infinite kind of learning in a unified way. Let us start with its definition.

Definition 4 (SL projection learning [3]) *Let S be a closed subspace in H, to which a target function belongs, and L be a complementary subspace of $S \cap \mathcal{N}(A)$ in S:*

$$S = L \dot{+} \{S \cap \mathcal{N}(A)\}. \tag{20}$$

Let P be a linear operator that becomes the projection operator onto L along $S \cap \mathcal{N}(A)$ when it is restricted to S. An operator X is called an SL projection learning operator and denoted by $A^{(\mathrm{SL})}$ if X, under the constraint

$$XAP_S = PP_S, \tag{21}$$

minimizes

$$J_n[X] = E_{\boldsymbol{n}} \|X\boldsymbol{n}\|^2. \tag{22}$$

Obtaining f_0 by an $A^{(\mathrm{SL})}$ is called SL projection learning. The totality of SL projection learnings for all S and L is called a family of projection learnings.

SL projection learning comes from the following idea. Projection learning, partial projection learning, and averaged projection learning have the following common properties:

(i) XA is a projection operator when A is restricted to a subspace of H.

(ii) Eq.(11) is minimized under the condition (i).

Definition 4 expresses these two properties as follows.

In projection learning, partial projection learning, and averaged projection learning, the target function f belongs to H, S, and $\overline{\mathcal{R}(R)}$, respectively. These subspaces are denoted by a closed subspace S in a unified way. In other words, we consider the case that f belongs to S. Let P be a linear operator that becomes a projection operator when it is restricted to S. Then, property (i) is described by

$$XAP_S = PP_S. \tag{23}$$

This is still different from Eq.(21), since the meaning of the operator P is not specified as is in Definition 4.

Eq.(23) has a solution X if and only if

$$\mathcal{N}(P) \supset S \cap \mathcal{N}(A), \tag{24}$$

which yields

$$S \cap \mathcal{N}(P) \supset S \cap \mathcal{N}(A). \tag{25}$$

Eq.(23) implies that the signal component XAf becomes 0 for f in $S \cap \mathcal{N}(P)$. Hence, we consider the smallest case of $S \cap \mathcal{N}(P)$ in Eq.(25):

$$S \cap \mathcal{N}(P) = S \cap \mathcal{N}(A). \tag{26}$$

Let L be a complementary subspace of $S \cap \mathcal{N}(A)$ in S. Then, P is defined as in Definition 4, and the property (i) is expressed by Eq.(21). Eq.(22) is a direct expression of the property (ii).

An SL projection learning operator has the following general form. Let A_s be an operator defined by

$$A_s = AP_S. \tag{27}$$

A_s is a sampling operator for functions in S. Since the target function belongs to S, the operator A_s plays an essential role rather than A. $\mathcal{R}(A_s)$ is the subspace which consists of all images of $f \in S$ under A. Hence, all elements in $\mathcal{R}(A_s)$ could be the signal component Af of the observed signal $\boldsymbol{y}$. That implies that noise vectors $\boldsymbol{n}$ in $\mathcal{R}(A_s)$ can not be distinguished from Af. We can distinguish $\boldsymbol{n}$ only outside of $\mathcal{R}(A_s)$ from Af. Therefore, when $\mathcal{R}(A_s)$ is the entire space $\mathbf{C}^M$, we can not suppress noise. In this paper, we discuss the case that $\mathcal{R}(A_s)$ is a proper subspace of $\mathbf{C}^M$.

Let us define a subspace S_t by

$$S_t = \mathcal{R}(A_s) + \mathcal{R}(Q), \tag{28}$$

where Q is the correlation matrix of the noise ensemble:

$$Q = E_{\boldsymbol{n}}(\boldsymbol{n} \otimes \overline{\boldsymbol{n}}). \tag{29}$$

$\mathcal{R}(Q)$ is the smallest subspace that includes all noise vectors $\boldsymbol{n}$ in the sense of mean square [14]. Hence, we take only $\boldsymbol{n}$ in $\mathcal{R}(Q)$ into account, hereafter. Since Af is a vector in $\mathcal{R}(A_s)$, $\boldsymbol{y} = Af + \boldsymbol{n}$ belongs to the subspace S_t. Let P_{S_t} be the orthogonal projection matrix onto S_t.

The subspace S_t has the following direct sum decomposition [3]:

$$S_t = \mathcal{R}(A_s) \dot{+} Q\mathcal{R}(A_s)^{\perp}. \tag{30}$$

Then, $\mathbf{C}^M$ can be decomposed as follows:

$$\begin{aligned} \mathbf{C}^M &= S_t \oplus S_t^{\perp} \\ &= \mathcal{R}(A_s) \dot{+} \{Q\mathcal{R}(A_s)^{\perp} \oplus S_t^{\perp}\}, \end{aligned} \tag{31}$$

where $\oplus$ denotes the orthogonal direct sum. Let P_t be the projection matrix onto $\mathcal{R}(A_s)$ along $Q\mathcal{R}(A_s)^{\perp} \oplus S_t^{\perp}$.

Theorem 1 ([3]) *An operator X is an SL projection learning operator if and only if*

$$XP_{S_t} = PA_s^{\dagger}P_t, \tag{32}$$

where $A_s^{\dagger}$ is the Moore-Penrose generalized inverse of A_s. A general form of $A^{(\mathrm{SL})}$ is given by

$$A^{(\mathrm{SL})} = PA_s^{\dagger}P_t + Y(I_M - P_{S_t}), \tag{33}$$

where I_M is the identity matrix on $\mathbf{C}^M$ and Y is an arbitrary operator. The minimum value, J_{n0}, of $J_n[X]$ is given by

$$J_{n0} = \langle PA_s^\dagger P_t Q, PA_s^\dagger \rangle, \tag{34}$$

where $\langle \cdot, \cdot \rangle$ is the Schmidt inner product of operators [10].

It follows from the definition of P_t that

$$P_t = P_t P_{S_t}.$$

Hence, Eq.(32) implies that $A^{(\mathrm{SL})}$ is an SL projection learning operator if and only if it maps a vector u in S_t into $PA_s^\dagger P_t u$ and that in $S_t^\perp$ into arbitrary functions in H. This is reflected in Eq.(33). In fact, since $I_M - P_{S_t}$ is the orthogonal projection matrix onto $S_t^\perp$, $A^{(\mathrm{SL})}$ transforms vectors in S_t and $S_t^\perp$ by the first term $PA_s^\dagger P_t$ and the second term Y of Eq.(33), respectively. Since $\boldsymbol{y}$ is a vector in S_t, the function f_{SL} obtained by SL projection learning is expressed as follows:

Corollary 1 ([3]) *f_{SL} is given by*

$$f_{\mathrm{SL}} = PA_s^\dagger P_t \boldsymbol{y}. \tag{35}$$

Let us define $\boldsymbol{y}_0$ by

$$\boldsymbol{y}_0 = P_t \boldsymbol{y}. \tag{36}$$

Then, Eq.(35) yields

$$f_{\mathrm{SL}} = PA_s^\dagger \boldsymbol{y}_0. \tag{37}$$

Transformations in Eqs.(36) and (37), i.e., P_t and $PA_s^\dagger$, are illustrated in Figure 1 with thick arrows. In order to understand them, the following corollary is important.

Corollary 2 *([3]) An operator X is an SL projection learning operator if and only if*

$$Xu = \begin{cases} PA_s^\dagger u & : u \in \mathcal{R}(A_s), \\ 0 & : u \in Q\mathcal{R}(A_s)^\perp. \end{cases} \tag{38}$$

The subspace S has the following direct sum decomposition [3]:

$$S = \mathcal{R}(A_s^*) \oplus \{S \cap \mathcal{N}(A)\}. \tag{39}$$

Hence, we have $P_S = A_s^\dagger A_s + P_{S \cap \mathcal{N}(A)}$. Therefore,

$$PP_S = P(A_s^\dagger A_s + P_{S \cap \mathcal{N}(A)}) = PA_s^\dagger A_s.$$

Then, Eq.(21) yields

$$XA_s = PA_s^\dagger A_s, \tag{40}$$

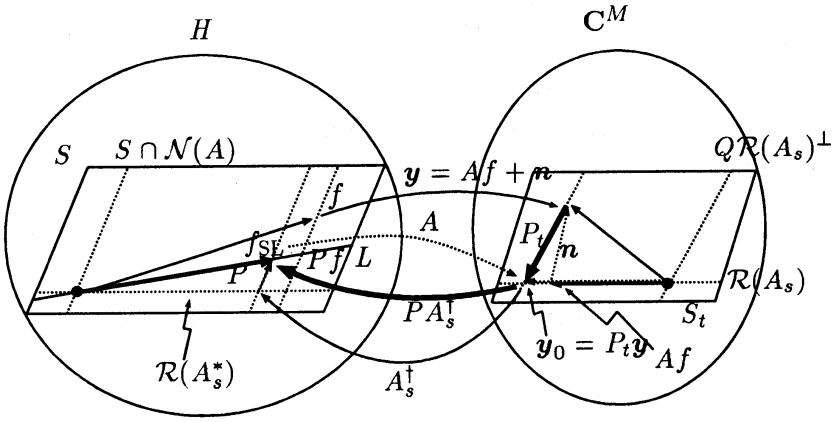


Figure 1: Mechanism of SL projection learning.

which implies that Eq.(21) is equivalent to the first equation of Eq.(38). Moreover, we can see from Eq.(20) that A is a bijection from L to $\mathcal{R}(A_s)$. Therefore, whether X is an SL projection learning operator, i.e., whether X minimizes Eq.(22) under the constraint of the first equation of Eq.(38), is determined from which complementary subspace of $\mathcal{R}(A_s)$ in S_t is mapped into zero. The second equation of Eq.(38) means that Eq.(22) is minimized by mapping $Q\mathcal{R}(A_s)^\perp$ into zero. In other words, $Q\mathcal{R}(A_s)^\perp$ is the subspace that contains the largest amount of noise vectors among complementary subspaces of $\mathcal{R}(A_s)$ in S_t. This noise suppression is achieved by the matrix P_t.

The operator A is a surjection from L to $\mathcal{R}(A_s)$. Hence, an original element in L which is mapped into $\boldsymbol{y}_0$ by A is unique. Since P and $P_{\mathcal{R}(A_s^*)}$ are the inverse transformations each other between L and $\mathcal{R}(A_s^*)$, and it holds that $P_{\mathcal{R}(A_s^*)} = A_s^\dagger A_s$, the operator A and $PA_s^\dagger$ are the inverse transformations each other between L and $\mathcal{R}(A_s)$. Therefore, Eq.(37) means that f_{SL} is the original element of $\boldsymbol{y}_0$.

The noise suppression of SL projection learning is evaluated in the function space H as we can see in Eq.(22). Note that it is realized in the vector space $\mathbf{C}^M$ by mapping vectors in $Q\mathcal{R}(A_s)^\perp$ into 0.

The general form of $A^{(\mathrm{SL})}$ in Eq.(33) can be simplified depending on the nature of noise. For example,

Theorem 2 ([3]) *If and only if*

$$Q\mathcal{R}(A_s)^\perp \subset \mathcal{R}(A_s)^\perp, \tag{41}$$

$A^{(\mathrm{SL})}$ *is expressed by*

$$A^{(\mathrm{SL})} = PA_s^\dagger + Y(I_M - P_{S_t}). \tag{42}$$

The first term $PA_s^\dagger P_t$ of the right-hand side of Eq.(33) becomes $PA_s^\dagger$ in Eq.(42). Since $\mathcal{N}(A_s^\dagger) = \mathcal{R}(A_s)^\perp$, elements in $\mathcal{R}(A_s)^\perp$ are mapped into 0 by $A_s^\dagger$. Hence, if Eq.(41) holds,

elements in $Q\mathcal{R}(A_s)^\perp$ are also mapped into 0 by $A_s^\dagger$. Therefore, the noise suppression of SL projection learning is realized by $A_s^\dagger$ without P_t.

For example, if variances of noise n_m are identically σ^2 and uncorrelated each other, then Q is given by $\sigma^2 I$, and the condition (41) holds.

SL projection learning becomes projection learning when

$$S = H, \quad L = \mathcal{R}(A^*). \tag{43}$$

In this case, $A_s = A$ and $P = P_{\mathcal{R}(A^*)}$. Then, Theorem 1 and Corollary 1 yield

Corollary 3 ([3]) *A general form of $A^{(\mathrm{P})}$ is given by*

$$A^{(\mathrm{P})} = A^\dagger P_t + Y(I_M - P_{S_t}). \tag{44}$$

A function, f_{P}, obtained by projection learning is given by

$$f_{\mathrm{P}} = A^\dagger P_t \boldsymbol{y}. \tag{45}$$

Note that, in this corollary, P_{S_t} is the orthogonal projection matrix onto the subspace $S_t = \mathcal{R}(A) + \mathcal{R}(Q)$, and P_t is the projection matrix onto $\mathcal{R}(A)$ along $Q\mathcal{R}(A)^\perp \oplus S_t^\perp$.

Furthermore, Theorem 2 reduces to

Corollary 4 *If and only if*

$$Q\mathcal{R}(A)^\perp \subset \mathcal{R}(A)^\perp, \tag{46}$$

$A^{(\mathrm{P})}$ is expressed by

$$A^{(\mathrm{P})} = A^\dagger + Y(I - P_{S_t}). \tag{47}$$

f_{P} is given by

$$f_{\mathrm{P}} = A^\dagger \boldsymbol{y}. \tag{48}$$

Eqs.(45) and (48) are easily calculated by a computer as follows. It holds that

$$A^\dagger = A^*(AA^*)^\dagger. \tag{49}$$

Because of Eq.(4), AA^* in Eq.(49) is a matrix whose i,j-th element is $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. It is denoted by K. Hence, the Moore-Penrose generalized inverse of the operator A is calculated by using that of the matrix K. As a result, we have the following algorithm for projection learning:

Algorithm 1 *1. Calculate the matrix K.*

2. Calculate the vector $\boldsymbol{z} = K^\dagger P_t \boldsymbol{y}$.

3. Obtain f_{P} in Eq.(45) by

$$f_{\mathrm{P}}(\boldsymbol{x}) = \sum_{m=1}^{M} z_m K(\boldsymbol{x}, \boldsymbol{x}_m), \tag{50}$$

where z_m is the m-th element of the vector $\boldsymbol{z}$.

Eq.(48) is also calculated by substituting the identity matrix I_M for P_t in the step 2.

Similarly, SL projection learning becomes partial projection learning when

$$L = \mathcal{R}(P_S A^*), \tag{51}$$

and averaged projection learning when

$$S = \overline{\mathcal{R}(R)}, \quad L = \mathcal{R}(RA^*). \tag{52}$$

Hence, for example, a general form of the partial projection learning operator $A^{(\mathrm{PTP})}$ and the averaged projection learning operator $A^{(\mathrm{AP})}$ are expressed by

$$A^{(\mathrm{PTP})} = A_s^\dagger P_t + Y(I - P_{S_t}), \tag{53}$$

$$A^{(\mathrm{AP})} = RA^*(ARA^*)^\dagger AA_s^\dagger P_t + Y(I - P_{S_t}), \tag{54}$$

respectively.

5 Minimum Variance Projection Learning

By using the theory of the family of projection learnings, a new kind of learning is introduced.

Definition 5 ([2]) *For a fixed subspace S, SL projection learning is called 'minimum variance projection learning' if the subspace L minimizes J_{n0} in Eq.(34).*

Due to noise in training examples, resultant functions for a fixed f are not unique in general. They are distributed around a function obtained from noiseless training examples. The smaller the variance of the distribution is, the more stable results can be obtained. The variance is expressed by J_{n0}. These discussions lead us to Definition 5.

Minimum variance projection learning can be characterized as follows:

Theorem 3 ([2]) *SL projection learning is minimum variance projection learning if and only if L satisfies*

$$L \supset L_N, \tag{55}$$

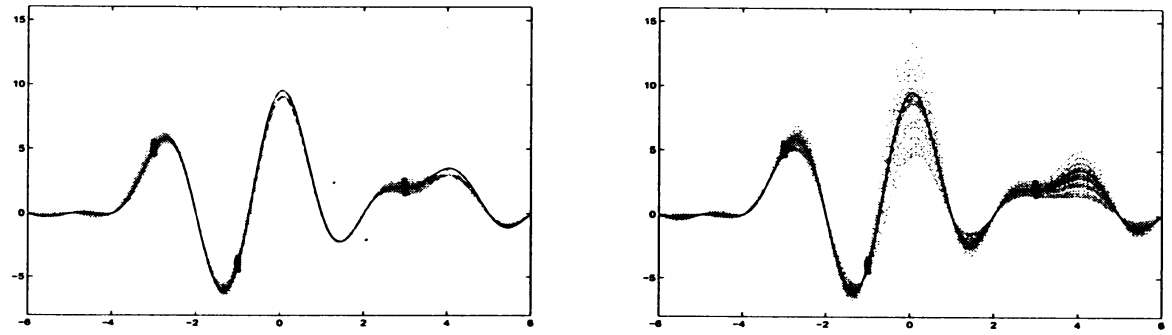
where L_N is a subspace defined as

$$L_N = \mathcal{R}(A_s^\dagger P_t Q). \tag{56}$$

In this case, J_{n0} is given by

$$J_{n0} = \langle A_s^\dagger P_t Q, A_s^\dagger \rangle. \tag{57}$$

Figure 2 (a) and (b) show functions obtained by minimum variance projection learning and non-minimum variance SL projection learning, respectively, from thirty sets of training examples. They are shown by dotted lines. In both figures, $f(x)$ and $(Pf)(x)$ are also shown by solid and dashed lines, respectively. These figures show that minimum variance projection learning provides more stable results than other SL projection learnings.



(a) (b)

Figure 2: Comparison of minimum variance projection learning (MVPL) and non-MVPL. (a) Resultant functions by MVPL from 30 sets of training examples. (b) Resultant functions by non-MVPL from the same training examples as in (a).

6 Conclusion

In this paper, we reviewed a functional analytic theory of supervised learning. First, we formulated the problem of supervised learning by using a reproducing kernel Hilbert space. Then, projection learning, partial projection learning, and averaged projection learning were introduced. Furthermore, these individual theories of learning were unified into a theory of the family of projection learnings. It treats an infinite kind of learning in a unified way. By using the theory, we introduced minimum variance projection learning.

Model selection and active learning are also important topics in supervised learning. The former is the problem of determining parameters in learning such as the subspace S, while the latter is that of designing the sample points for optimal generalization. They can also be discussed in this framework (see e.g. [11, 12]).

References

[1] Stefan Bergman, The Kernel Function and Conformal Mapping, American Mathematical Society, Rhode Island, 1970.

[2] Akira Hirabayashi and Hidemitsu Ogawa, "Projection learning of the minimum variance type," In Proc. 1999 Int. Conf. on Neural Information Processing, vol.3, pp.1172–1177, Perth, Australia, November 1999.

[3] Akira Hirabayashi and Hidemitsu Ogawa, "A family of projection learnings," The Transactions of the Institute of Electronics, Information and Communication Engineers D-II, vol.J83-D-II, no.2, pp.754–767, February 2000 (in Japanese). Its short English version titled "A class of learning for optimal generalization" appeared in

Proc. 1999 Int. Joint Conf. on Neural Networks, no.246 (CD-ROM), Washington D.C., USA, July 1999.

[4] Hidemitsu Ogawa, "Projection filter regularization of ill-conditioned problem," In Proc. of SPIE, vol.808, Inverse Problems in Optics, pp.189–196, March 1987.

[5] Hidemitsu Ogawa, "Inverse problem and neural networks," In Proc. IEICE 2nd Karuizawa Workshop on Circuits and Systems, pp.262–268, May 1989 (in Japanese).

[6] Hidemitsu Ogawa, "Neural network learning, generalization and over-learning," Proc. ICIIPS'92, Int. Conf. on Intelligent Information Processing & System, Beijing, China, Supplemental volume, pp.1–6, Oct. 1992.

[7] Hidemitsu Ogawa, "Neural networks and generalization ability," The Institute of Electronics, Information and Communication Engineers Technical Report, no.NC95-8, pp.57–64, May 1995 (in Japanese).

[8] Hidemitsu Ogawa and Shoji Hara, "Properties of partial projection filter," The Transactions of the Institute of Electronics, Information and Communication Engineers A, vol.J71-A, no.2, pp.527–533, February 1988 (in Japanese).

[9] Hidemitsu Ogawa and Kazutaka Yamasaki, "Generalization and over-learning of neural networks," The Institute of Electronics, Information and Communication Engineers Technical Report, no.NC91-75, pp.77–84, January 1992 (in Japanese).

[10] Robert Schatten, Norm Ideals of Completely Continuous Operators, Springer-Verlag, New York, 1970.

[11] Masashi Sugiyama and Hidemitsu Ogawa, "Incremental active learning for optimal generalization," Neural Computation, vol.12, no.12, pp.2909–2940, December 2000.

[12] Masashi Sugiyama and Hidemitsu Ogawa, "Subspace information criterion for model selection," Neural Computation, vol.13, no.8, pp.1863–1889, August 2001.

[13] Sethu Vijayakumar, Masashi Sugiyama, and Hidemitsu Ogawa, "Training data selection for optimal generalization with noise variance reduction in neural networks," In Proc. WIRN Vietri-98, The 10-th Italian Workshop on Neural Nets, pp.153–166, Salerno, Italy, May 1998.

[14] Yukihiko Yamashita and Hidemitsu Ogawa, "Properties of averaged projection filter for image restoration," The Transactions of the Institute of Electronics, Information and Communication Engineers D-II, vol.J74-D-II, no.2, pp.142–149, February 1991 (in Japanese). Its English translation appeared in Systems and Computers in Japan, Scripta Technica, Inc. USA, vol.23, no.1, pp.69–78, April 1992.